HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

Bachelorarbeit

Herr
Toni Trodler

Analyse von Transkriptionsfaktoren und Rückschlüsse auf evolutionäre Zusammenhänge

Mittweida, 2013

Fakultät Mathematik, Naturwissenschaften,
Informatik

BACHELORARBEIT

Analyse von Transkriptionsfaktoren und Rückschlüsse auf evolutionäre Zusammenhänge

Autor:
Herr

Toni Trodler

Studiengang:
Biotechnologie/Bioinformatik

Seminargruppe:
BI 10 W1 - B

Erstprüfer:
Prof. Dr. rer. nat. Dirk Labudde

Zweitprüfer:
Dipl. Ing. Daniel Stockmann

Einreichung:
Mittweida, 26.08.2013

Verteidigung/Bewertung:
Mittweida, 2013

Bibliographische Beschreibung:

Trodler, Toni: Analyse von Transkriptionsfaktoren und Rückschlüsse auf evolutionäre Zusammenhänge. - 2013. - Seitenzahl Verzeichnisse 5, Seitenzahl des Inhaltes 37, Seitenzahl der Anhänge 1 S. Mittweida, Hochschule Mittweida, Fakultät: Mathematik, Naturwissenschaften, Informatik, Bachelorarbeit, 2013

Englischer Titel

Investigation of transcription factors and conclusion to evolutionary connections.

Kurzbeschreibung:

Diese Bachelorarbeit mit dem genannten Thema beschäftigt sich mit der Analyse von Transkriptionsfaktoren. Das sind DNA - bindende Proteine, welche die Stärke der Transkription beeinflussen können und somit positive als auch negative Auswirkungen auf die Genexpression haben. Die meisten solcher Bindestellen befinden sich im Bereich -500 bis 100 bp relativ zur Startstelle der Transkription. Außerdem findet ein Vergleich dieser regulatorischen Regionen zwischen den Spezies Mensch, Maus und Ratte statt. Dieses Verfahren wird phylogenetisches Footprinting genannt.

Danksagung

In allererster Linie möchte ich mich bei Prof. Dr. Dirk Labudde, meinem Betreuer bedanken, der mein Interesse an diesem Thema bereits im 5 Semester meines Studiums geweckt hat. Dieses Thema entspricht genau meinem Interessensbereich im großen Zweig der Biotechnologie und Bioinformatik. Er hat mich hervorragend in der Zeit meines Praktikums und der nachfolgenden Bachelorarbeit betreut und sich Zeit genommen um aufgekommene Fragen zu klären. Die Zusammenarbeit mit ihm hat mir sehr gut gefallen.
Weiterhin möchte ich mich bei Tilman Sauer bedanken, der mir mit seiner Doktorarbeit sehr geholfen hat und ich somit ein besseres Verständnis für mein Thema bekommen habe.
Zuletzt danke ich noch meiner Familie und meinen Freunden, die sich Zeit genommen haben, um etwas Abwechslung in den Arbeitstag zu bringen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AS	Aminosäure(n)
bzw.	beziehungsweise
bp	Basenpaare
ca.	circa
DNA	desoxyribonucleic acid
HLH	Helix-Loop-Helix
HTH	Helix-Turn-Helix
ID	Identifikator
IUPAC	International Union of Pure and Applied Chemistry
mind.	mindestens
Mio.	Millionen
Mrd.	Milliarden
mRNA	messenger ribonucleic acid
n.a.	nicht angegeben
PSSM	Positions-spezifische-scoring-Matrix
TBP	TATA - Box binding Protein
TF	Transkriptionsfaktor
TFBS	Transkriptionsfaktor - Bindestelle
TSS	Transkriptionsstartstelle
z.B.	zum Beispiel

1 Einleitung

Die Genomgröße verschiedener eukaryontischer Organismen beträgt zwischen 10 Millionen und 3 Milliarden Basenpaaren. Durch die immer schnelleren und kostengünstigeren Sequenziermöglichkeiten sind bis heute viele wichtige Organismen vollständig sequenziert. Diese Tatsache schafft für die Bioinformatik weitreichende Möglichkeiten die Masse an Sequenzen mit verschiedenen Tools zu untersuchen. Denn die Fülle an Sequenzinformation die dahinter steckt, ist kaum aufgeklärt. Die molekularbiologischen Mechanismen, welche die Entwicklung von komplexen Organismen steuern sind nahezu unbekannt. Der Mensch besitzt circa 22'000 Gene, welche jeweils ein Protein transkribieren. Fakt ist jedoch, dass 95% des gesamten Genoms gar nicht transkribiert werden und dass rund 30% davon allein für die Kontrolle der Genexpression verantwortlich sind. Daher ist es von größter Bedeutung die Sequenzinformationen aufzuklären und ein Verständnis für die Kontrollelemente eines Transkriptionsapparates zu bekommen.

Dafür muss geklärt sein, wie die Gene in Eukaryonten Aufgebaut sind. Die regulatorischen Bereiche befinden sich meist vor dem Gen was transkribiert wird. Es handelt sich um Enhancer, also Transkriptionsfaktor - Bindestellen (TFBS), die genauer untersucht werden sollen.

Moderne Methoden wie das DNA - Footprinting oder das EMSA (Electrophoretic Mobility Shift Assay) erlauben den Nachweis der Interaktion von Transkriptionsfaktoren (Proteinen) mit den entsprechenden TFBS auf der DNA. Der Nachteil ist hierbei, dass die analytische Bestimmung im Labor sehr teuer ist, daher ist es das Ziel der Bioinformatik die reine Sequenz auf solche TFBS zu untersuchen, da dies kostensparender ist.

Die Herausforderung der Bioinformatik ist in diesem Gebiet die Charakterisierung der regulatorischen Elemente. Die Schwierigkeit besteht darin, dass diese meist sehr kurz von ihrer Sequenzlänge sind, mehrere Kilobasen von der eigentlichen Startstelle entfernt sein können und zudem noch sehr variabel innerhalb der verschiedenen Organismen sind.

Das weitreichende Verständnis der zellulären Abläufe kann im Laufe der Zeit ein wichtiger Schritt zum bekämpfen von genetisch bedingten Krankheiten sein. [1]

2 Biologische Grundlagen

Bei den folgenden Punkten soll auf biologische Grundlagen eingegangen werden, deren Verständnis notwendig ist, um die angewandten Methoden besser zu verstehen.

2.1 Aufbau der DNA

Bei der DNA handelt es sich um ein Biomolekül, das sich bei eukaryontischen Organismen in jeder Zelle innerhalb des Zellkerns befindet. Sie ist dicht verpackt in den Nukleosomen, welche aus den Histonproteinen H1 - H4 bestehen. Die Nukleosomen sind wiederrum gebündelt und bilden im ganzen ein komplettes Chromosom.
In der DNA ist die gesamte genetische Information eines Organismus enthalten. Sie ist ein lineares Biopolymer, welches aus Nukleotiden besteht. Dieses setzen sich aus einem Phosphatrest, Desoxiribose (Zucker) und einer von 4 verschiedenen stickstoffhaltigen Basen zusammen. Die Phosphatgruppe und der Zuckerbaustein formen gemeinsam das Rückgrat eines Stranges. Wie in Abbildung 1 zu sehen ist, liegen sich immer zwei Basen gegenüber. Diese sind jeweils an einem Kohlenstoffatom des Zuckermoleküls gebunden. Im inneren gehen die Basenpaare gemeinsam eine Wasserstoffbrückenbindung ein, wobei immer Adenin und Thymin eine Zweifachbindung und Guanin und Cytosin eine Dreifachbindung ausbilden.
Nummeriert man den 5 - Fach - Zucker der Kohlenstoffatome nach durch, so befindet sich an Position 1' die Base, am 3' - Ende die OH - Gruppe und am 5' - Ende der Phosphatrest. Diese Eigenschaft führt dazu, dass man den beiden DNA - Strängen eine eindeutige Richtung zuordnen kann, nämlich 5' → 3' oder 3' → 5'. Die negativ geladenen Phosphatgruppen verleihen dem gesamten Doppelstrang eine negative Ladung. Anhand dieser wichtigen Eigenschaft kann man die DNA in einem elektrischen Feld wandern lassen. Weiterhin findet ein stärkerer Zusammenhalt der Stränge statt, wenn Guanin mit Cytosin gepaart ist, da drei Wasserstoffbrücken für einen stabileren Halt sorgen als zwei. Daher ist die Bestimmung des GC - Gehaltes von Bedeutung.[3]

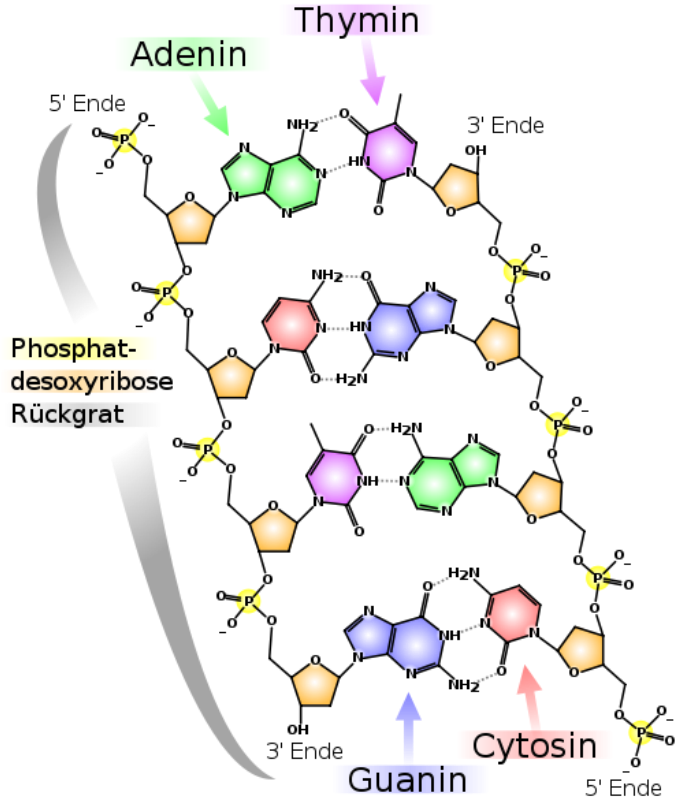


Abbildung 1: Aufbau der Doppelhelix [URL-2]

2.2 Gene und ihre zugehörigen Promotorsequenzen

Gene sind Bereiche auf der DNA-Sequenz, die für ein Protein codieren. Die Anzahl der Gene auf der Erbinformation reicht bei Eukaryonten von 2000 bis zu 45000. In Tabelle 1 ist eine Übersicht über Genomgröße, Anzahl der Gene und Chromosomenzahl der in dieser Arbeit verwendeten Organismen Mensch, Maus und Ratte gegeben.

Tabelle 1: genetische Kenngrößen von Mensch, Maus und Ratte [URL-3]

	Genomgröße	Gene	Chromosomen
Mensch	3,0 Mrd.	25000	46
Maus	2,6 Mrd.	30000	40
Ratte	2,75 Mrd.	n.a	42

Würden Gene den gesamten Platz der menschlichen DNA einnehmen, so hätte jedes im Schnitt eine Größe von 120'000 bp. Dies ist aber nicht der Fall, da ein Großteil der Erbinformation nicht-codierend ist. Bei Hefen, ist dieser Anteil ca. 50%, bei Säugetieren schon ca. 85%. Hierbei lässt sich erkennen, je komplexer ein Organismus

ist, umso größer wird der nicht-codierende Bereich. Es gibt also Bereiche, die unserem Wissensstand zufolge keine Bedeutung haben. Diese Annahme ist aber nicht sicher.
In Abbildung 2 erkennt man, dass solche nicht-codierenden Bereiche sich auch innerhalb eines Gens befinden können, sogenannte Introns. Diese werden nach der Transkription aus der prä-mRNA herausgeschnitten (Splicing), die Exons miteinander verbunden und es entsteht eine fertige mRNA. Diese wird nach den Regeln des genetischen Codes, an den Ribosomen, in eine Polypeptidkette umgeschrieben (Translation), sodass ein Protein entsteht[3]

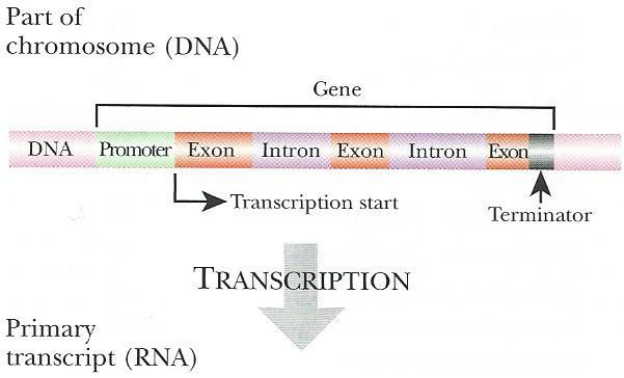


Abbildung 2: Gene befinden sich auf der DNA [3]

Der typische Aufbau eines eukaryontischen Gens ist in Abbildung 3 gezeigt. Vor dem Gen, welches ein Protein codiert, befindet sich eine Promotorregion, die zur Regulation der Transkription von Bedeutung ist. An ihr können DNA-bindende Proteine andocken, sogenannte Transkriptionsfaktoren. Der Promotor befindet sich nahe der Startseite der Transkription und enthält seinerseits typische Elemente, die zur Initiation der Transkription erforderlich und in vielen eukaryontischen Polymerase II-Genen enthalten sind. Auf einem Promotor liegen verschiedene Elemente. Die Initiator-Box, die TATA-Box und je nach Gen verschiedene Upstream-Elemente. Die Initiator-Box ist der Bereich, wo sich die TSS befindet, oft als +1 gekennzeichnet. Ungefähr 25 bp Upstream gelegen ist die TATA-Box, eine AT-Reiche Sequenz, die von TBP gebunden wird. Weiter Stromaufwärts befindet sich oftmals eine GC-Box, sowie eine CAAT-Box. Je nachdem wie weit ein Element von der TSS liegt, werden drei Arten von Promotoren unterschieden. Der Kernpromotor umschließt den Bereich von der TSS bis -35 bp. Weiterhin liegt der proximale Promotor -35 bp bis -250 bp Stromaufwärts. Der distale

Promotor, in Abbildung 3 auch Enhancer-Region genannt, umfasst alle weiteren zum Promotor gehörenden TFBS und kann noch viele tausende bp entfernt von der TSS liegen.

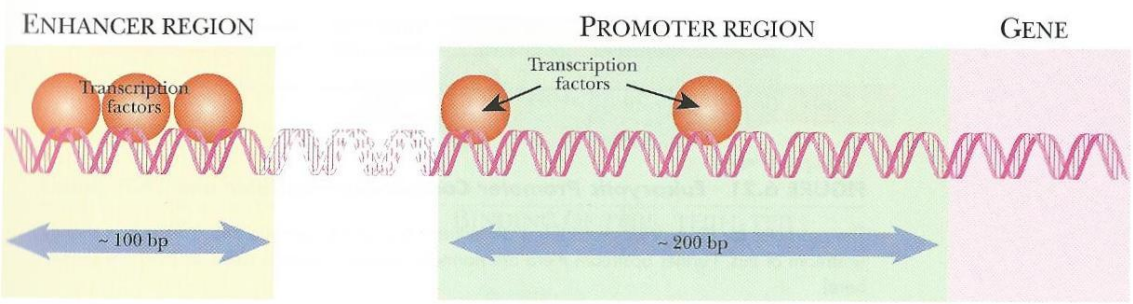


Abbildung 3: Promotorregion befindet sich vor dem Gen [3]

Promotoren unterscheiden sich anhand der Kombination der cis-regulatorischen Elemente. Man spricht dann auch von einer kombinatorischen Genexpression, die je nach Zelltyp unterschiedlich ist. Die Bereiche zwischen den TFBS haben keine genetische Bedeutung, können aber Spacer-Elemente sein, die eine Schlaufenbildung bewirken und damit die Funktion haben andere Elemente in bestimmte Positionen zu bringen. Insgesamt sind Promotorsequenzen sehr variabel in ihrer Sequenz und dadurch schwer zu charakterisieren, da z.B. regulatorische Elemente mit mehreren Kilobasen Abstand zur TSS vorkommen können. [1][3]

2.3 Transkriptionsfaktoren und Transkriptionsfaktorbindestellen

Transkriptionsfaktoren sind DNA-bindende-Proteine, welche die Transkriptionsrate positiv als auch negativ beeinflussen können. Sie lassen sich in die allgemeinen- und spezifischen Transkriptionsfaktoren unterteilen. Die Allgemeinen TF werden für alle Gene benötigt, an denen eine Polymerase beteiligt ist und bezeichnen sich mit TFI, TFII oder TFIII gefolgt von einem Individuellen Buchstaben. Die Zahlen I bis III bezeichnen die an der Transkription beteiligte RNA-Polymerase I, II oder III. Die allgemeinen Transkriptionsfaktoren haben drei wichtige Aufgaben zu erfüllen. Zum ersten sind sie notwendig dafür die RNA Polymerase II richtig an den Promotor zu positionieren. Weiterhin trennen sie die beiden DNA Stränge, damit die Transkription starten kann und im späteren Verlauf eine Freisetzung der RNA Polymerase während der Transkription ermöglichen. [4]

In Tabelle 2 ist eine Übersicht über die Funktion der allgemeinen Transkriptionsfaktoren gegeben.

Tabelle 2: Die allgemeinen Transkriptionsfaktoren mit ihren Funktionen [3]

Transkriptionsfaktor	Funktion
TBP	Bindet TATA - Box, Untereinheit von TFIID
TFIID	Beinhaltet TBP, erkennt Polymerase II - Promotor
TFIIA	Bindet Upstream der TATA - Box, benötigt für die Bindung von RNA - Polymerase II an den Promotor
TFIIB	Bindet downstream der TATA - Box, benötigt für die Bindung von RNA - Polymerase II an den Promotor
TFIIF	Hilft RNA - Polymerase II an den Promotor zu binden
TFIIE	Benötigt für die Ablösung der RNA - Polymerase II vom Promotor und zur anschließenden Elongation
TFIIH	Phosphoryliert den RNA - Polymerase II - Schwanz und wird während der Elongation beibehalten
TFIIJ	Benötigt für die Ablösung der RNA - Polymerase II vom Promotor und zur anschließenden Elongation

Allgemeine Transkriptionsfaktoren wirken im Bereich des Kernpromotors, da sie nur Funktionen bezüglich der RNA-Polymerase erfüllen. Das bedeutet, dass sie keinen regulatorischen Einfluss ausüben. In Abbildung 4 ist dies verdeutlicht. Zu erkennen ist die RNA-Polymerase mit ihren allgemeinen TF, sowie die TATA-Box an der TBP gebunden ist. Stromaufwärts befinden sich dann die spezifischen TF. Diese wirken nicht an jedem Gen, sondern nur an bestimmten. Dies ist abhängig von äußeren Signalen wie z.B.: Hormone, Stoffwechselverbindungen usw. In Eukaryonten ist es oft der Fall, dass ein solches Signal, das ein TF anschaltet, eine ganze Signalkaskade oder auch Pathway durchläuft. Daraus lässt sich schließen, dass spezifische Transkriptionsfaktoren ihrerseits durch äußere Umweltsignale reguliert werden.

Solche spezifischen TF erkennen eine bestimmte DNA-Sequenz im proximalen und distalen Promotorbereich und binden an diese. Außerdem stehen sie entweder direkt oder indirekt in Kontakt mit dem Transkriptionsapparat. Die Schwierigkeit solche

Bindestellen auf der DNA zu erkennen ergibt sich daraus, dass diese sehr kurz sind und zudem noch variabel in ihrer Abfolge. [1][4]

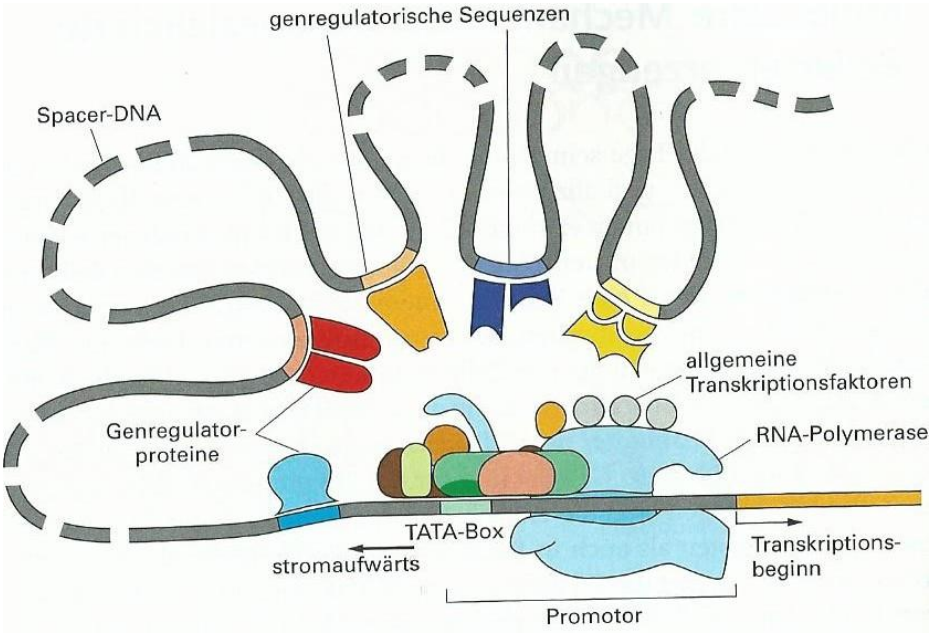


Abbildung 4: komplexer eukaryontischer Transkriptionsapparat [4]

2.3.1 Arten von Transkriptionsfaktoren

Fast alle TF lagern sich in die große Furche der DNA ein, da sie somit besseren Kontakt zu den Basen erhalten. Dort finden molekulare Wechselwirkungen statt. "Das Protein bildet Wasserstoffbrückenbindungen, Ionenbindungen und hydrophobe Wechselwirkungen mit den Basen, ohne dabei im Normalfall die Wasserstoffbrückenbindungen zu stören, welche die Basen zusammenhalten."[1]

Spezifische Transkriptionsfaktoren lassen sich nochmal in verschiedene Arten unterteilen. Man unterscheidet Helix-Turn-Helix-, Helix-Loop-Helix-, Leucin-Zipper- und Zink-Finger-Motive. In Abbildung 5 sind die ersten zwei genannten Motive gezeigt. Beide bestehen aus zwei α-Helices verbunden durch einen Turn (Teilbild A) bzw. einen Loop (Teilbild B). Bei dem HTH-Motiv setzt sich die zweite Helix (Vom N-Terminus aus gezählt) in die große DNA-Furche hinein. Im Gegensatz dazu ist das HLH-Motiv an einer Dimerisierung beteiligt. Die DNA-Bindung gelangt durch einen basischen Abschnitt von Aminosäuren, meist auf der N-Terminalen Seite. [3][4]

[1] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter (2005): *Lehrbuch der Molekularen Zellbiologie*. 3. Auflage: WILEY-VCH Verlag: S. 288

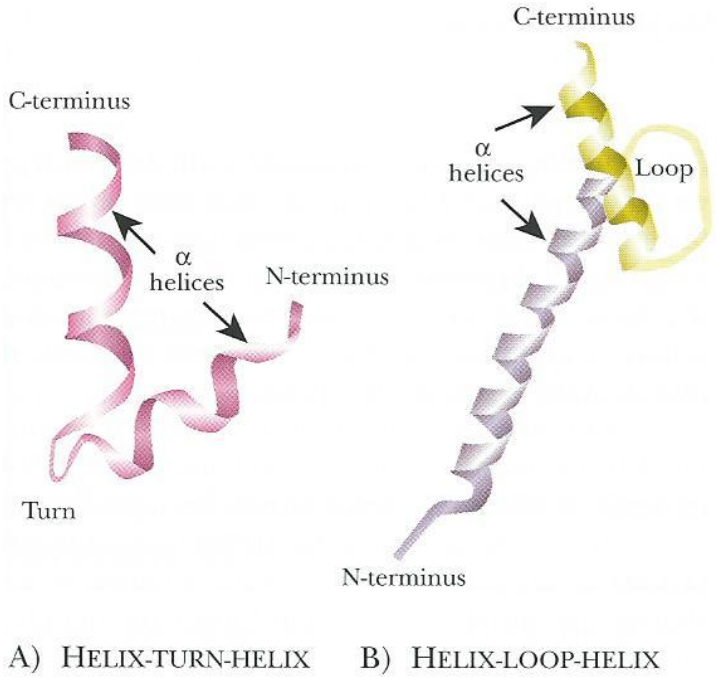


Abbildung 5: HTH und HLH - Bindemotiv von Transkriptionsfaktoren [3]

Das Leucin-Zipper Bindemotiv zeigt sich bei vielen eukaryontischen TF wie z.B.: Fos, Jun oder Myc-Proteinen. Diese TF sind beteiligt in der Kontrolle der Zellteilung und Karzinogenese. In Abbildung 6 ist ein solches Motiv dargestellt. Es besteht aus einer α-Helix mit Leucin-Resten, die alle sieben Aminosäuren vorkommen. Außerdem sind die Aminosäuren zwischen den Leucinen meist hydrophob. Da 3,6 Aminosäuren pro Windung vorkommen, bilden die hydrophoben Reste einen Strang auf einer Seite der Helix. Wie man in Abbildung 6 erkennen kann, binden zwei antiparallel verlaufende α-Helices durch die Hydrophoben Leucin-Stränge aneinander und formen somit eine Reißverschluss-Struktur. Die Bindung an die große Furche der DNA erfolgt durch die basische Region des Proteins. [3]

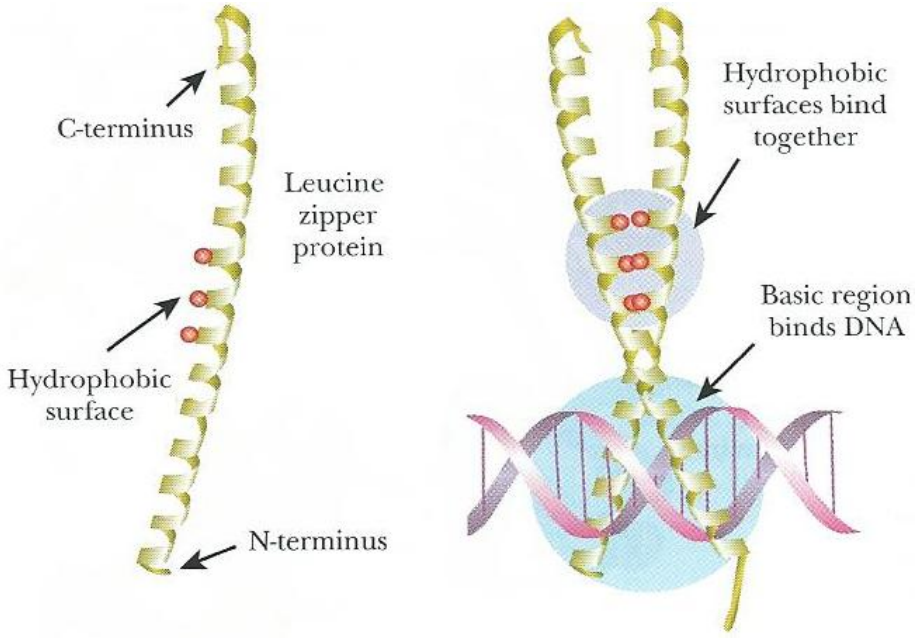


Abbildung 6: Leucin-Zipper Bindemotiv [3]

Ein letztes Motiv, dass häufig als TF vorkommt ist der Zink-Finger. Dieser besteht aus einem zentralen Zink-Atom mit einem Abschnitt von 25-30 AS-Resten, die um das Atom herum angeordnet sind. Das Zink-Atom ist an die beiden Schwefelatome vom Cystein (In Abbildung 7 mit Symbol C gekennzeichnet), sowie an zwei Stickstoff-Atome vom Histidin (H) gebunden. [3]

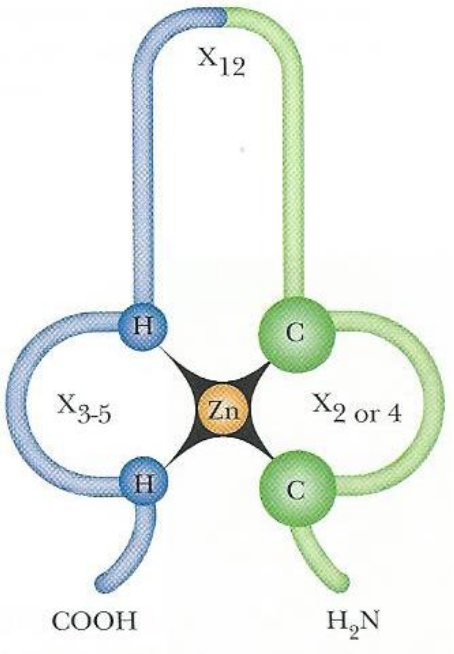


Abbildung 7: Zink-Finger Bindemotiv [3]

2.3.2 Lokalisierung von Transkriptionsfaktoren

Deletionsanalyse von Upstream - Regionen

Die Upstream gelegene Region eines Strukturgens beinhaltet verschiedene Bindungsstellen für Regulatorproteine, wie z.B.: die Promotorregion. Um solche stellen zu finden, gibt es verschiedene Methoden.

Eine Möglichkeit ist die Schrittweise Entfernung von Regulator Regionen vom 5' Ende. Damals setzte man dies mit Restriktionsenzymen um. Heute geschieht dies durch PCR. Mit Hilfe von modernem Primer design kann man sehr genau die Regionen flankieren und somit verschiedene Längen dieser Upstream - Region erzeugen. Diese Sequenz Segmente verschiedener Längen werden in Organismen fusioniert und die Expression der Strukturgene wird untersucht. In Abbildung 8 ist dieses Vorgehen in Teilbild I dargestellt.

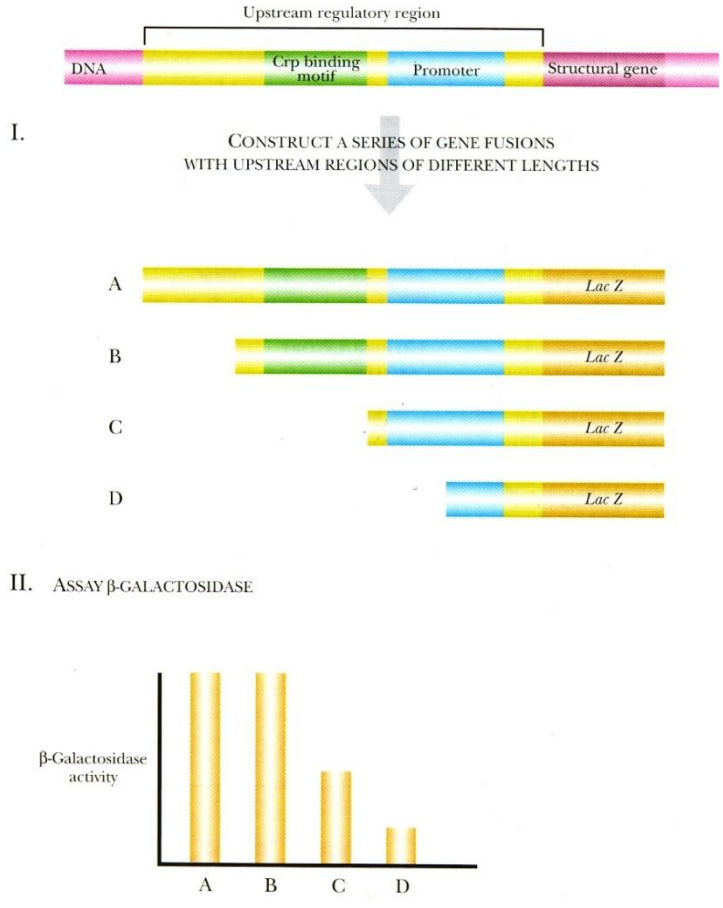


Abbildung 8: Schrittweise Deletion der Regulatorregion und Messen der β-Galactosidase-Aktivität [3]

In diesem Beispiel besitzt die komplette Upstream Region (A) eine hohe Aktivität. Das entfernen eines Teilstückes (B) bewirkt eine vernachlässigbaren Effekt, was darauf schließen lässt, dass keine wichtigen Sequenzabschnitte in dieser Region liegen. Wird jedoch die Crp Region entfernt, sinkt die Aktivität um die Hälfte (C). Daraus kann man schlussfolgern, dass die Crp - Bindungsstelle die Genexpression stark beeinflusst. Entfernt man sogar die Hälfte des Promotors (D), ist die ß - Galactosidase Aktivität verschwindend gering. Die Ergebnisse bestätigen, dass die Crp- und die Promotor - Region, die Expression der Strukturgene beeinflussen. Zunächst weißt dieses Experiment nach, ob es bestimmte Bindungsmotive gibt, welche die Genexpression beeinflussen. Es zeigt aber nicht, ob ein Protein wirklich an diese Sequenz bindet.

EMSA - Eletrophoretic Mobility Shift Assay

Um zu überprüfen, ob ein Regulatorprotein Einfluss auf die Genexpression hat, testet man ob dieses Protein an die DNA bindet. Voraussetzung hierfür ist, dass man aufgereinigte DNA, sowie das aufgereinigtes Protein getrennt vorliegen hat. Die Geschwindigkeit eines DNA Fragmentes in einem Elektrophorese Gel ist verändert, wenn ein Protein daran bindet. Zunächst schneidet man mit Hilfe von Restriktionsenzymen den DNA Abschnitt in mehrere kleinere Fragmente. Nun werden 2 Ansätze von diesen Proben gemacht. In einen wird das Regulator Protein zugegeben, in den anderen nicht. Diese 2 Ansätze werden elektrophoretisch aufgetrennt und untersucht. Im Falle einer erfolgreichen Bindung des Regulator Proteins mit der DNA wird diese Bande, im Vergleich zur Bande ohne zugegebenen Regulator Protein, etwas kürzer gelaufen sein. [3]

2.4 Konsensus - Sequenzen und PSSMs

Zum Auffinden von TFBS auf der DNA verwendet man oft bereits bekannte TFBS. Dabei wird mit der Sequenz einer bekannten TFBS auf einer neuen Sequenz gesucht um diese dort wiederzufinden. Dieses Verfahren nennt sich "String-matching". Da Diese Bindestellen sehr variabel sein können, ist diese Methode unpraktisch und man benötigt genauere Eingabewerte zur Suche. Mehr Informationsgehalt bieten zum Beispiel Konsensussequenzen. Diese zeigen Unterschiede in einer Reihe von TFBS auf, sodass beim Konsensus-matching schon mehr positive Ergebnisse als beim String-matching

erhalten werden. In Tabelle 3 sind die Regeln für eine IUPAC-Konsensussequenz aufgelistet.

Tabelle 3: IUPAC-Zeichen für Konsensussequenzen [1]

Nukleotid	IUPAC-Zeichen
A oder G	R
C oder T	Y
A oder C	M
G oder T	K
C oder G	S
A oder T	W
A, C oder T	H
C, G oder T	B
A, C oder G	V
A, G oder T	D
A, C, G oder T	N

Um eine solche Konsensussequenz zu bestimmen, werden zunächst alle TFBS aufgelistet, aligniert und durchnummeriert. Schließlich wird die Häufigkeit eines Nukleotids an jeder Position bestimmt. Wenn die relative Häufigkeit einer Base größer als 50% ist und die der zweithäufigste Base kleiner oder gleich 25%, gilt das Nukleotid als konserviert. Haben zwei Basen zusammen eine Häufigkeit von mind. 75%, gilt der IUPAC-Code für die zwei Nukleotide, aber nur falls Regel eins nicht zutrifft.

Den größten Informationsgehalt besitzen Positions-spezifische Scoring-Matrizen, da sie jede mögliche Position optimal nach der Anzahl der Nukleotide wichten. In Tabelle 4 ist die Erstellung einer PSSM dargestellt. Es sind sechs verschiedene TFBSs mit einer Länge von 9 bp aufgelistet. Die daraus folgende PSSM hat demnach die Länge neun und eine Breite von vier (Die Basen A, C, G, T). In den Spalten ist für jede Base die Häufigkeit an der entsprechenden Position angegeben. In der letzten Zeile ist zudem noch die IUPAC-Konsensussequenz dargestellt.

Tabelle 4: mehrere TFBS werden zu einer PSSM bzw. Konsensussequenz umgeformt [1]

Position		1	2	3	4	5	6	7	8	9
TFBSs		G	T	G	A	C	T	C	A	G
		A	T	G	A	C	T	C	A	G
		A	T	G	A	C	A	T	C	A
		C	T	G	A	C	T	C	A	T
		A	T	G	A	C	T	A	A	C
		G	T	G	A	C	G	A	A	A
PSSM	A	3	0	0	6	0	1	2	5	2
	C	1	0	0	0	6	0	3	1	1
	G	2	0	6	0	0	1	0	0	2
	T	0	6	0	0	0	4	1	0	1
IUPAC Konsensus-Seq		R	T	G	A	C	T	M	A	N

"PSSMs nutzen einen größeren Anteil der Sequenzinformation als Konsensussequenzen und sind daher akkurater in der Vorhersage neuer potentieller TFBSs."[2] Ein negativer Aspekt ist, dass die einzelnen Positionen unabhängig betrachtet werden und somit Wechselbeziehungen zwischen Positionen verloren gehen. [1]

Das Ziel ist es mit PSSM eine DNA-Sequenzen nach Mustern abzusuchen und zwar so, dass die Summenverteilung der Basen optimal der DNA-Sequenz entsprechen. Dazu wird die Matrix so umgeformt, dass man mit ihr Rechnen kann. Aus den Häufigkeiten der Basen an Position X entsteht dann über Log-Odd-Scores ein neuer Zahlenwert. Tabelle 5 zeigt diese Umformung. [URL-4]

Tabelle 5: Umformen einer PSSM mit Hilfe von Log-Odds [URL-4]

A	5	0	1	0	0
C	0	2	2	4	0
G	0	3	1	0	4
T	0	0	1	1	1

$$\xrightarrow{\text{Log}\left(\frac{f(b,i)+s(n)}{p(b)}\right)}$$

A	1,6	-1,7	-0,2	-1,7	-1,7
C	-1,7	0,5	0,5	1,3	-1,7
G	-1,7	1	-0,2	-1,7	1,3
T	-1,7	-1,7	-0,2	-0,2	-0,2

Um mit der neu Berechneten PSSM Bindestellen für Transkriptionsfaktoren zu finden geht man wie folgt vor. Man startet mit dem Fenster bei Position 1 auf der Eingabesequenz und addiert für jede Position den Score der entsprechenden Base auf. Danach rückt man das Fenster immer um eine Position weiter. Wenn "n" die Länge der Sequenz ist und "k" die Größe des Fensters, dann erhält man auf diese Weise n - k + 1 Score-Werte. In Abbildung 9 ist als Beispiel somit der höchste Score von

[2] Sauer, Tilman 2006 *Evaluierung des phylogenetischen Footprintings zur verbesserten Vorhersage von Transkriptionsfaktor-Bindestellen*. 148 Seiten, Göttingen, Georg - August Universität Göttingen, Mathematisch - Naturwissenschaftliche Fakultät: S.20

Abs_score = 13.4 entstanden. Damit gilt diese Stelle als positiver Treffer für eine TFBS des Transkriptionsfaktors Sp1. [URL-4]

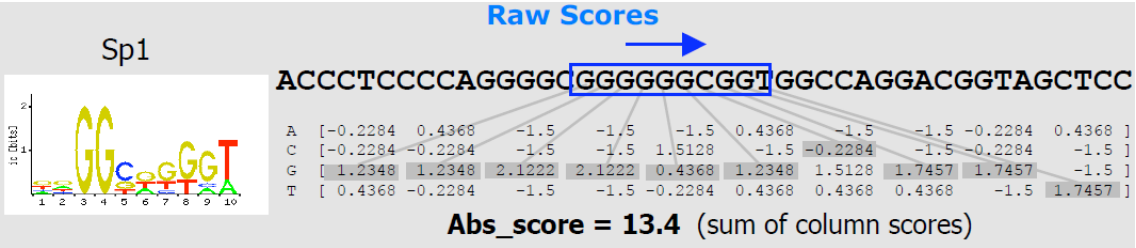


Abbildung 9: Mit PSSM Schrittweise den Score eines Fensters berechnen [URL-4]

2.5 Paarweises Sequenz-Alignment

Bei einem paarweisen Sequenz-Alignment werden zwei Sequenzen miteinander verglichen um herauszufinden, ob diese durch Mutation oder Selektion aus einem Vorfahren entstanden sind. Dies sind natürliche Prozesse, da Substitutionen, Insertionen oder Deletionen während der Replikation der DNA auftreten können. Dabei ist es so, dass bestimmte Mutationen häufiger auftreten als andere. Nachdem die Sequenzen Aligniert wurden, kann man feststellen, ob es sich um eine Verwandtschaft oder Zufall der Sequenzen handelt. Dabei ordnet man die Sequenzen so an, dass jede Position der ersten Sequenz einer Position der zweiten entspricht. Es sind auch Lücken für den Fall der Deletion oder Insertion erlaubt, welche mit einem "-" Symbol bezeichnet werden. Stimmen zwei Basen nicht überein, ist das eine Fehlpaarung bzw. Substitution. In Abbildung 10 ist ein paarweises Sequenzalignment einer Promotorregion zwischen Mensch und Maus dargestellt.

```
Mensch  gttttatgacaa--agaaaattttc-tg
        ||||   ||||||   ||||.||||||  ||
Maus    gttt--tgacaaagagaacattttcatg
```

Abbildung 10: Paarweises Sequenz-Alignment Mensch/Maus [URL-5]

Die senkrechten Verbindungen zwischen 2 Basen stehen für eine Übereinstimmung, der " . " für eine Substitution und das " - " für eingebaute Lücken.

Um das erhaltene Alignment zu bewerten, benötigt man eine Scorefunktion. Oft verwendet man sogenannte Substitutions-Matrizen, dabei werden Übereinstimmende Basen positiv bewertet, Substitutionen eher negativ und Lücken erhalten eine "gap penalty" (Strafe). Der Score wird dann durch Addition für alle Werte des

Residuen-Paares errechnet, wobei angenommen wird, dass Mutationen zufällig auftreten. Ein optimales Alignment erhält man, wenn man den Score maximiert. Für unsere Promotorsequenzen bietet es sich immer an ein globales Alignment durchzuführen. Dabei wird jeder Base der einen Sequenz eine der anderen Sequenz oder auch einer Lücke zugeordnet. Beim lokalen Alignment passiert das nur für Teilsequenzen. [1]

2.6 Phylogenetisches Footprinting

Beim Phylogenetischen Footprinting handelt es sich um ein Verfahren zur Identifizierung von Transkriptionsfaktor - Bindestellen innerhalb der nicht codierenden Bereiche der DNA. Also genauer dem regulatorischen Bereich vor dem Protein (Promotorregion). Diese Bereiche werden mit anderen Spezies verglichen, indem sie miteinander Aligniert werden. Die Annahme auf der das Verfahren beruht ist, dass Bereiche, die für die Regulation der Genexpression verantwortlich sind einem höheren evolutionären Druck unterliegen, als nicht funktionelle Bereiche. Der Grund dafür ist, dass Mutationen in TFBS, die deren Funktion ausschalten durch Selektion entfernt werden. Daraus schließt man, dass regulatorische Bereiche im Laufe der Evolution weniger Mutationen aufweisen und somit stärker konserviert sind.

Wichtig ist, dass man die richtigen Sequenzen miteinander Vergleicht. Man benötigt orthologe Sequenzen um zu bestimmen, welche Bereiche mehr Mutationen haben als andere. Ortholog bedeutet hierbei, dass sich diese Sequenzen über unzählige Generationen aus einer Vorläufersequenz herausgebildet und ihre Funktion behalten haben. Ein Alignment zeigt dann häufig eine höhere Übereinstimmung im Bereich der funktionellen Sequenzen. [1]

Zum Sequenzvergleich sollten die Spezies einen bestmöglichen Abstand haben. Sind sie zu Nah Verwandt ist nicht klar, ob die Sequenzen aufgrund des evolutionären Drucks oder aufgrund der gemeinsamen Abstammung konserviert sind. Zu weit entfernte Organismen bieten kein Aussagekräftiges Alignment.

Da heutzutage viele eukaryontische Organismen vollständig sequenziert sind, wird dieses Verfahren sehr häufig angewandt. Bei der Anwendung des phylogenetischen Footprintings beim Menschen hat sich herausgestellt, dass ein Vergleich mit der Maus die besten Ergebnisse liefert. Das liegt daran, dass diese beiden Organismen vor ca. 80 Mio. Jahren einen gemeinsamen Vorfahren besaßen. Sie besitzen den bestmöglichen

Abstand um das Verfahren darauf anzuwenden und möglichst viele positive Treffer zu erzielen.

Um dieses Verfahren selbstständig anzuwenden, wurde die Technik anhand der Disertation von Tilman Sauer selbstständig erarbeitet und dann Angewendet. In Abbildung 11 ist in einer kurzen Übersicht der allgemeine Ablauf des phylogenetischen Footprintings dargestellt. [1]

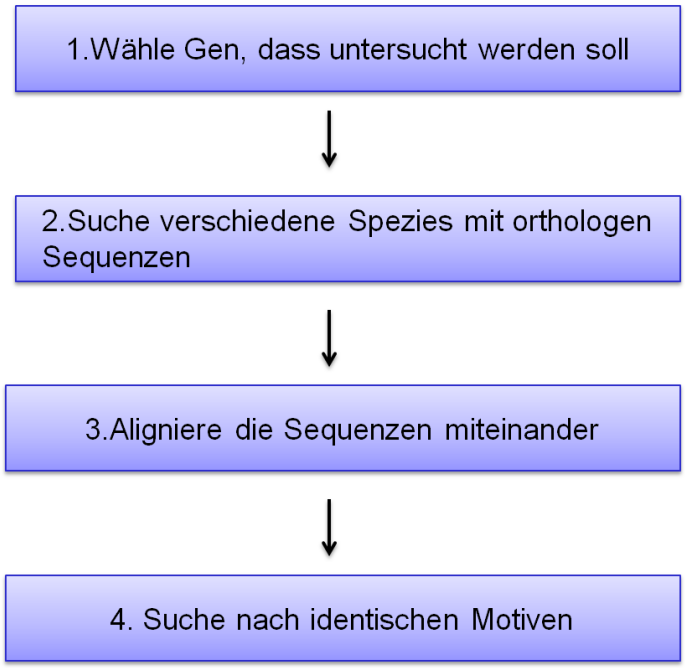


Abbildung 11: Vorgehen bei einem Phylogenetischen Footprinting

3 Verwendete Datenbanken und Programme

In diesem Kapitel werden kurz die Verwendeten, frei zugänglichen Datenbanken und Programme beschrieben.

3.1 EPD - Eukaryotic Promoter Database

Die Eukaryontische Promotor Datenbank (EPD) ist eine annotierte, nicht redundante Sammlung von eukaryontischen Polymerase II - Promotoren für welche die TSS experimentell bestimmt wurde. Der Zugang zu Promotorsequenzen ist durch den Punkt "Standard search" möglich. Die Annotation für jeden Eintrag umfasst eine Beschreibung der TSS, Referenzen zu anderen Datenbanken, sowie bibliografische Einträge. Diese Datenbank bietet eine schnelle Suche nach entsprechenden Promotorsequenzen zweier Spezies, um diese anschließend in anderen Programmen weiter zu analysieren. Die Datenbank umfasst insgesamt 25988 Promotorsequenzen für den Menschen, 9773 für die Maus und 11389 für D. melanogaster.
Dabei wird im Eingabefenster das offizielle Symbol für das Gen eingetragen und anschließend gesucht. Als Ergebnis erhält man verschiedene Promotoren von den oben genannten Spezies. Unter "View" lassen sich diese anzeigen. Danach gibt man die obere und untere Grenze der Länge bezogen auf die TSS an und bestätigt mit dem "Download" Button. Nun wird die Sequenz in einem neuen Fester angezeigt und kann kopiert werden. [URL-6]

3.2 GeneCards®

GeneCards ist eine Datenbank, die genomische Informationen zu allen bekannten und bereits annotierten menschlichen Genen bietet. Darüber hinaus besitzt sie ein Sammlung an Informationen über Transkriptomik, Genetik, Proteomik sowie funktionelle Krankheiten aus allen bekannten Quellen. Die Suchfunktion für Begriffe (z.B. Gene) ermittelt nach Eingabe einen Score, der anzeigt, welcher Eintrag dem Sucheintrag am nächsten entspricht. Transkriptionsfaktoren oder Gene besitzen teilweise mehrere unterschiedliche Schreibweisen. Die GeneCard-Datenbank zeigt alle Symbole für ein entsprechenden TF oder Gen an, unter dem es auch bekannt ist. Damit konnten größtmögliche Verwechselungen vermieden werden. Zum Beispiel ist der

Interleukin 2-Promotor unter dem Kürzel "IL-2" oder "TCGF" zu finden. Weiterhin werden sämtliche ID's für andere Datenbanken angegeben z.B.: Entrez Gene, Ensembl, UniProtKB und weitere. Solche Vernetzungen innerhalb von Datenbanken wurden fast überall gefunden. Damit ist eine eindeutige Identifizierung möglich und es besteht keine Verwechselungsgefahr zwischen Genen.
Durch den Verweis auf andere Datenbanken, folgte z.B. nach Eingabe eines Gens eine Weiterleitung zur
SABiosciences Regulatory transcription factor binding sites - Datenbank, bei der TF mit deren Bindestellen angezeigt wurden. [URL-7]

3.3 UCSC Genome Browser

Der UCSC Genome Browser ist eine Datenbank der University of California, welche Zugriff auf Genom-Sequenzen verschiedener Säugetiere bietet. Der Browser ist für graphische Darstellungen optimiert und wurde als web-basierendes Tool auf einer MySQL-Datenbank angepasst. In Abbildung 12 ist die Oberfläche des Genom Browsers für eine Genregion dargestellt. Durch Eingabefelder kann man verschiedenste Informationen zusätzlich hinzufügen bzw. abschalten. In Abbildung 12 sind TFBS unter dem Punkt "Regulatory elements from ORegAnno" zu finden. Durch Anklicken erhält man weitere Informationen dazu. [URL-8]

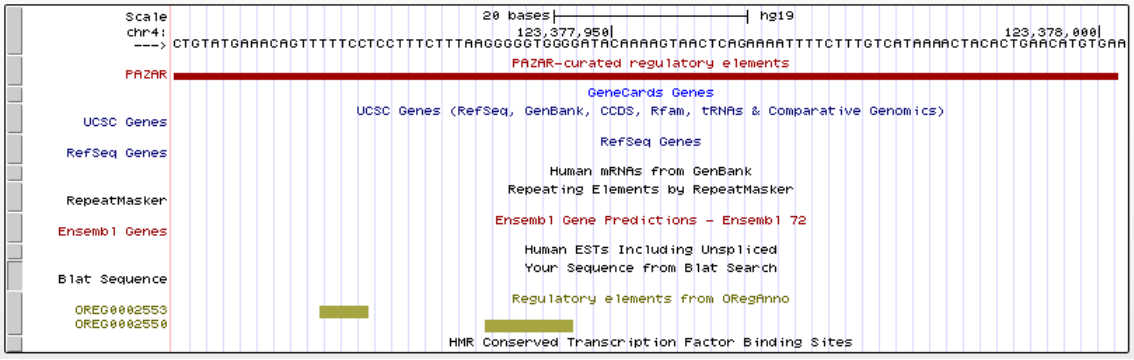


Abbildung 12: UCSC Genome Browser Oberfläche [URL-8]

3.4 cisRED

cisRED ist eine Datenbank und enthält Informationen über konservierte cis-regulatorische Elemente, welche eindeutig identifiziert und bewertet wurden. Auf cisRED befinden sich eine Vielzahl von Motiven aus ca. 7500 Sequenzsätzen, welche

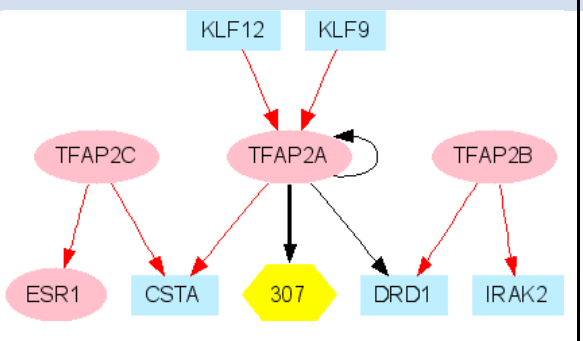
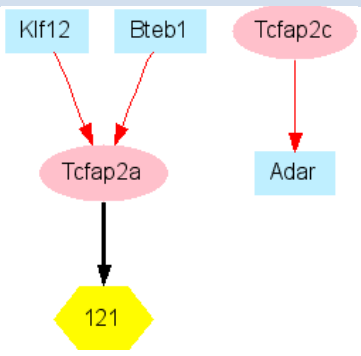
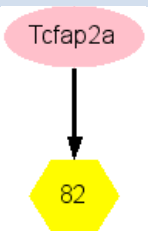
menschliche Promotoren mit durchschnittlich sechs homologen Sequenzen aus anderen Säugetieren enthalten. Ähnliche Motive wurden zusammengefasst und die Sequenzen haben eine Länge von -1500 bp bis 100 bp relativ zur TSS. Diese Datenbank ist frei verfügbar. In dieser Arbeit diente sie dazu nach Eingabe einer Gen ID potentielle Transkriptionsfaktoren zu finden. [1]

3.5 TRED - Transcriptional Regulatory Element Database

TRED ist eine weitere Datenbank mit einem für diese Arbeit wichtigen Tool. Sie wurde damals angelegt um cis- und trans-Regulatorische Elemente von Säugetieren zu integrieren und einen Zusammenhang beider zu schaffen. TRED bietet eine eigene Promotor-Annotation für Mensch, Maus und Ratte. Die Datenbank bietet Informationen zur Bindung von TF sowie deren Regulation.

Eine wichtige Übersicht bietet der Menüpunkt "Gene Regulatory Networks". In diesem sind die 36 TF-Familien von Mensch, Maus und Ratte angegeben. In Tabelle 6 ist das Gen-Regulatorische-Netzwerk für eine TF-Familie (AP2) gezeigt. Die roten Ellipsen sind IDs von Transkriptionsfaktoren und in den blauen Rechtecken befinden sich die Bezeichnungen für die Gene. Ein roter Pfeil weist darauf hin, dass eine TF-Gen-Interaktion erwiesen ist. Bei einem schwarzen Pfeil vermutet man dies nur. Die Zahlen in den gelben Hexagons beschreiben die Anzahl an Gen-Cluster. Die Gene sind separat aufgelistet, da sie sonst nicht in die Übersicht passen würden.

Tabelle 6: Gen-Regulator-Netzwerk für die AP2 TF-Familie [URL-13]

Mensch	Maus	Ratte
		

Die komplexität des Netzwerkes nimmt von Mensch, Maus bis hin zur Ratte ab. In der Bioinformatik beschäftigten sich viele mit der Erstellung solcher TF-Gen-Netzwerke.

Ein wichtiges Tool dieser Datenbank ist die "Sequence Matrix Search". Dieses Programm fordert als Eingabe-Wert die Promotorsequenz im FASTA-Format. Danach wird die Matrix gewählt mit der man die Sequenz nach Motiven untersuchen möchte. Ein großer Vorteil hierbei ist, dass dieses Tool direkt auf die JASPAR Datenbank mit all seinen Bindemotiven für Transkriptionsfaktoren zugreifen kann und somit die PSSM des entsprechenden TF selbstständig lädt. Eine weitere Möglichkeit ist es, die PSSM selbst hinzuzufügen. Um nicht jeden Treffer angezeigt zu bekommen, der einen Score größer als 0 hat, kann man in dem Fenster "Cutoff score" eine untere Grenze festlegen, ab der ein Treffer angezeigt wird. [URL-13]

3.6 ConSite

ConSite ist ein Web-Tool mit dem man ein phylogenetisches Footprinting durchführen kann. Man hat die Möglichkeit ein orthologes Paar oder eine einzelne Sequenz zu untersuchen. Dabei gibt man in Eingabefenstern die Sequenzen als FASTA-Format ein und beschriftet sie. Nach betätigen des Proceed-Buttons wählt man die TF aus, nach denen gesucht werden soll. Im Ergebnisfenster werden dann die Sequenzen mit ihrer entsprechenden Konserviertheit an den Nukleotid-Positionen angezeigt. Positive Hits für TF werden an der Sequenzposition angezeigt. Im Nachhinein hat man die Möglichkeit sich das Ergebnis unterschiedlich Darstellen zu lassen, z.B. als Alignment, als Tabelle oder das Profil für die Konservierung. [URL-9]

4 Methoden

Im folgenden Kapitel wird darauf eingegangen, was genau mit den Datenbanken gesucht und anschließend mit den im Kapitel 3 genannten Tools Analysiert wurde.

4.1 Finden von orthologen Sequenzpaaren

Zunächst wurde durch Internetrecherchen nach Genen gesucht, die bereits ausführlich untersucht sind. Von diesen konnte dann mit Hilfe der Eukaryotic Promoter Database (EPD) die Promotorsequenz gesucht werden. Parallel dazu sucht die Datenbank selbstständig nach orthologen Sequenzpaaren. In dieser Arbeit wird nur der Promotorbereich -700 bis +100 bp relativ zur TSS untersucht, da sich dort wie man in Abbildung 13 erkennen kann, die meisten Bindestellen für Transkriptionsfaktoren befinden.

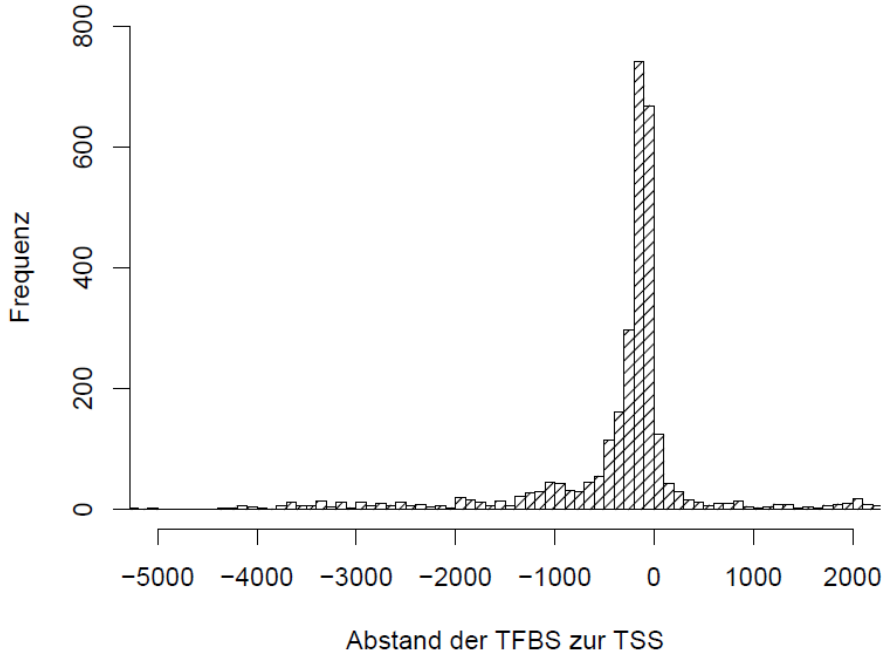


Abbildung 13: Reichweite von TFBS zur TSS [1]

Besitzt man orthologe Sequenzen von Mensch und Maus für ein bestimmtes Gen, werden diese zur Kontrolle paarweise Aligniert. Eine höhere Übereinstimmung von mehr als 70% weist darauf hin, dass die zwei Sequenzpaare korrekt sind.

4.2 Analyse der Sequenzen in ConSite

Man hat die Möglichkeit auch einzelne Sequenzen in ConSite zu untersuchen. Obwohl PSSMs relativ genau in ihrer Score-Wertung bei Sequenzen sind, werden unzählige Treffer für TFBS gefunden, selbst bei einer hohen "Cutoff-Schwelle". In Abbildung 14 ist eine Single-Sequenz Analyse gezeigt. Die hellblaue Linie unterhalb stellt die gesamte Eingabesequenz dar. Die positiven Treffer für TFBS des IL2-Promotor sind in dunkelblau darüber mit ihrer Gen ID gekennzeichnet.

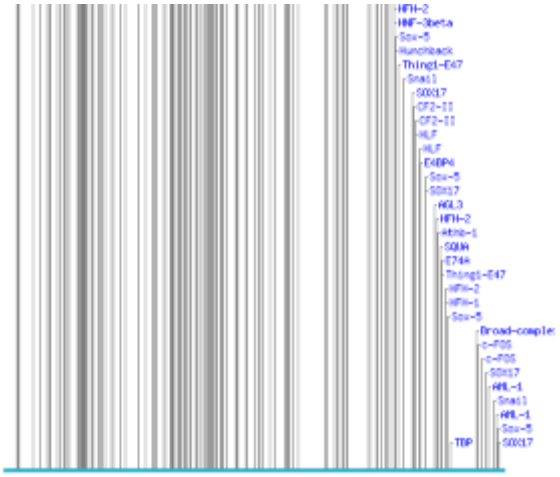


Abbildung 14: Treffer für TFBS [URL-9]

Diese unzähligen Treffer werden jedoch minimiert, wenn man das phylogenetische Footprinting darauf anwendet. Startet man ConSite nochmals mit zwei Sequenzen, nämlich die des Menschen und der Maus, werden die Treffer enorm reduziert. In Abbildung 15 wurde ebenfalls der IL2-Promotor untersucht, jedoch von Mensch und Maus. Die Treffer wurden ungefähr von hundert auf fünfundzwanzig reduziert. Wie in Kapitel 2.6 angesprochen, sind regulatorische Bereiche stärker Konserviert als nicht regulatorische. Genau diese Annahme kann in Abbildung 15 im unteren Diagramm nachgewiesen werden. Es zeigt die Konservierung der gesamten Eingabesequenz von Mensch und Maus an. Bereiche die über 92% konserviert sind werden als potentielle Regulatorsequenzen identifiziert. Sollte die Matrix-Suche dann einen positiven Treffer erzeugen gilt diese Stelle als TFBS für den entsprechenden TF.

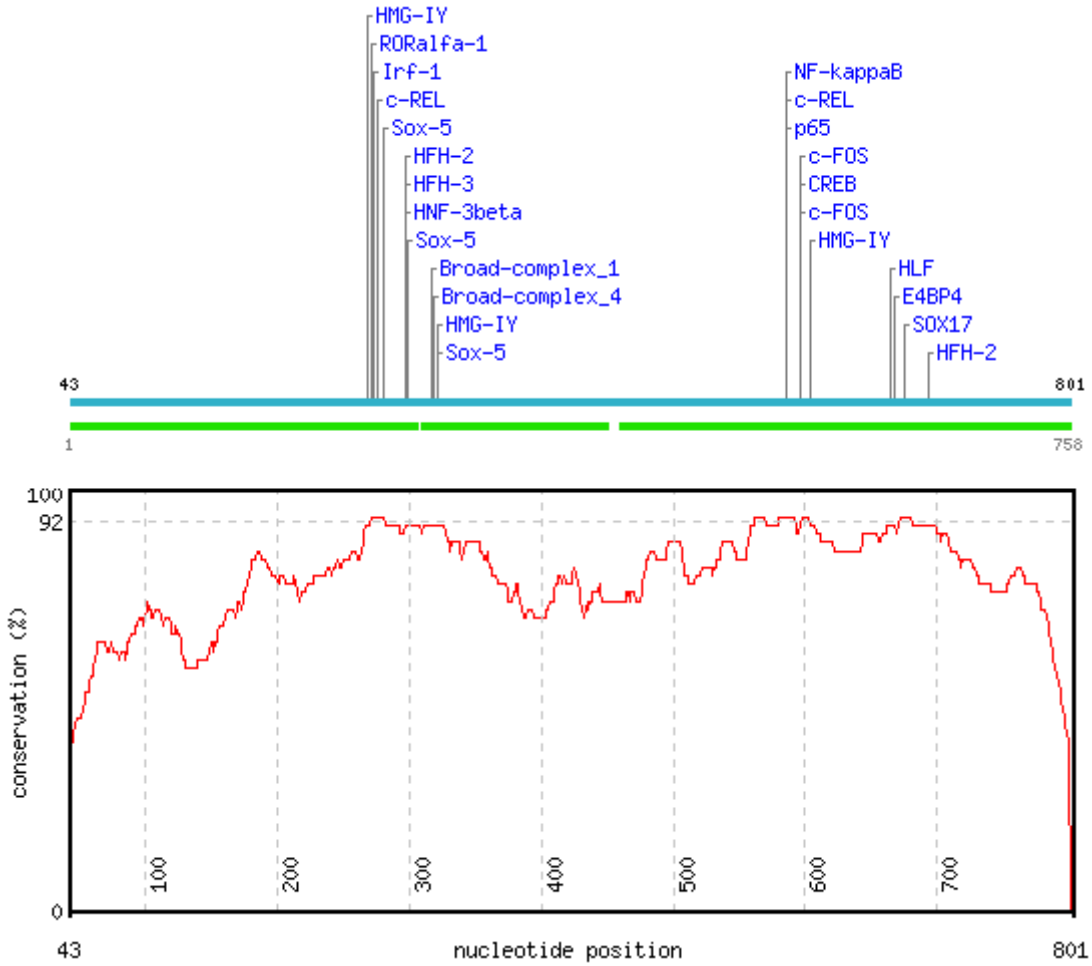


Abbildung 15: Konserviertheits-Profil des IL2 Promotors Mensch/ Maus [URL-9]

4.3 Vergleich mit anderen Datenbanken

Zum Vergleich, ob die Treffer echt positiv sind, wurde die Promotorsequenz mit anderen Datenbanken abgeglichen. TRED bietet eine grafische Darstellung von regulatorischen Netzwerken, cisRED enthält Informationen zu Transkriptionsfaktoren, die mit dem Eingabe-Gen interagieren und SABiosciences zeigt alle TF und einen ungefähren Ort der Bindestelle, sowie die Bindestelle selbst.

5 Ergebnisse

Im folgenden Kapitel werden die Ergebnisse für die verschiedenen Promotorregionen der entsprechenden Gene beschrieben.

5.1 IL10-Promotor

Die Promotor-ID IL10 steht für Interleukin 10 und erfüllt viele Funktionen innerhalb der Regulation des Immunsystems. [URL-7]

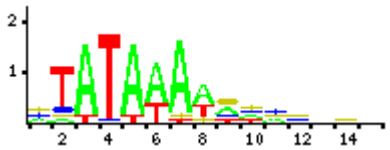
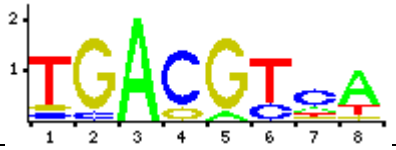
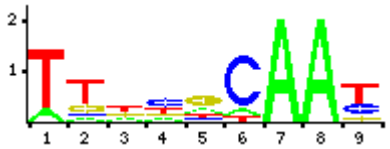
In Tabelle 7 ist eine Matrix-Suche auf der TRED-Datenbank gezeigt. Die markierten Sequenzteile kennzeichnen eine Bindestelle für TF mit einem relativ hohen Score. Dieser ist unter dem Namen des Transkriptionsfaktors mit dem Symbol "S" angegeben.

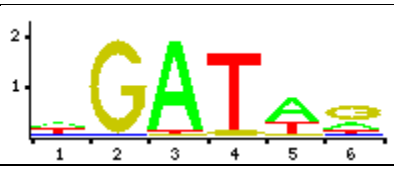
Tabelle 7: IL10-Promotor mit TFBS für die Matrixsuche [URL-10]

Menschlicher IL10-Promotor

>IL10:chr1:203589224 [-700..100](-) [human, Homo sapiens]

```
AGTTGGGGTGGGGGACAGCTGAAGAGGTGGAAACATGTGCCTGAGAATCC      50
TAATGAAATCGGGGTAAAGGAGCCTGGAACACATCCTGTGACCCCGCCTG     100
TACTGTAGGAAGCCAGTCTCTGGAAAGTAAAATGGAAGGGCTGCTTGGGA     150
ACTTTGAGGATATTTAGCCCACCCCCTCATTTTTACTTGGGGAAACTAAG     200
GCCCAGAGACCTAAGGTGACTGCCTAAGTTAGCAAGGAGAAGTCTTGGGT     250
ATTCATCCCAGGTTGGGGGGACCCAATTATTTCTCAATCCCATTGTATTC     300
TGGAATGGGCAATTTGTCCACGTCACTGTGACCTAGGAACACGCGAATGA     350
GAACCCACAGCTGAGGGCCTCTGCGCACAGAACAGCTGTTCTCCCCAGGA     400
AATCAACTTTTTTTAATTGAGAAGCTAAAAAATTATTCTAAGAGAGGTAG     450
CCCATCCTAAAAATAGCTGTAATGCAGAAGTTCATGTTCAACCAATCATT     500
TTTGCTTACGATGCAAAAATTGAAAACTAAGTTTATTAGAGAGGTTAGAG     550
AAGGAGGAGCTCTAAGCAGAAAAAATCCTGTGCCGGGAAACCTTGATTGT     600
GGCTTTTTAATGAATGAAGAGGCCTCCCTGAGCTTACAATATAAAAGGGG     650
GACAGAGAGGTGAAGGTCTACACATCAGGGGCTTGCTCTTGCAAAACCAA     700
ACCACAAGACAGACTTGCAAAAGAAGGCATGCACAGCTCAGCACTGCTCT     750
GTTGCCTGGTCCTCCTGACTGGGGTGAGGGCCAGCCCAGGCCAGGGCACC     800
C                                                      801
```

TATA-Box S = 9,17	
CREB S = 2,56	
cEBP S = 7,73	

GATA-3 S = 2,90	

Untersuchungen des Promotors bestätigen, dass die gefundenen TFBS tatsächlich echt-positiv sind und bereits nachgewiesen wurden. In Abbildung 16 sind alle experimentell Nachgewiesenen TFBS angegeben.

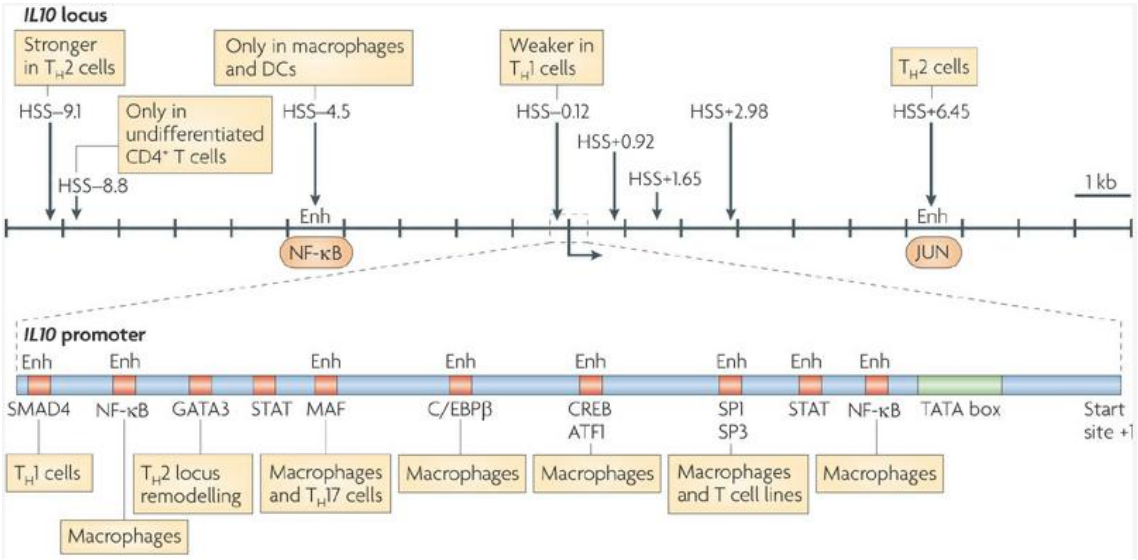


Abbildung 16: TFBS des Interleukin-10 Promotors [URL-11]

5.2 ACTA-1 Promotor

Der ACTA-1 Promotor gehört zum Gen, welches Actin alpha 1 exprimiert. Dieses Protein wird im Skeletmuskel exprimiert und ist eines von sechs verschiedenen Aktin-isoformen. Aktine sind hoch konservierte Proteine, die eine Rolle bei der Zellbeweglichkeit, Zellstruktur und des Schutzes spielt. [URL-7]

Ein phylogenetische Footprintanalyse mit ConSite ergab zwei positive Treffer. In Abbildung 17 sind die zwei Hits für die Transkriptionsfaktoren SRF und AGL3 gezeigt.

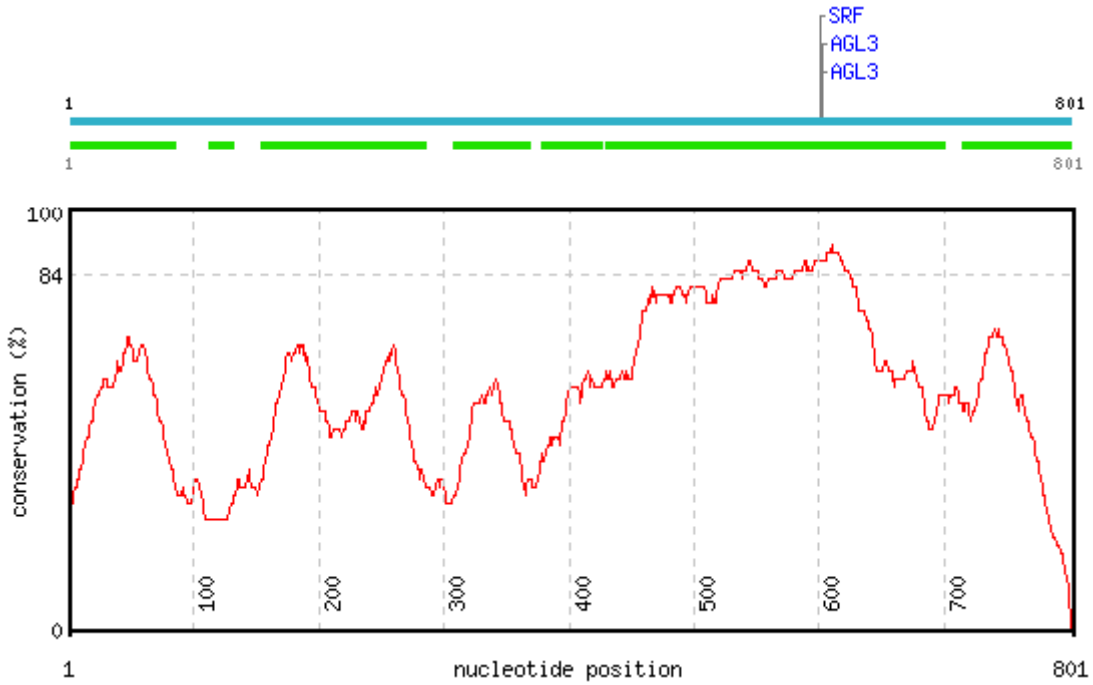
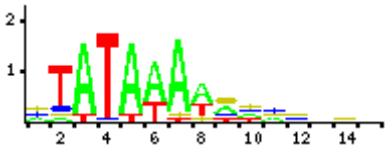
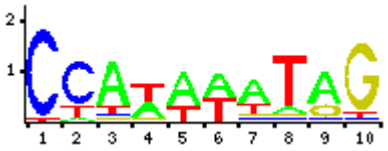
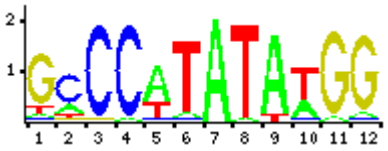


Abbildung 17: Phylogenetisches Footprinting des ACTA-1 Promotors [URL-9]

Die anschließende Matrixsuche mit TRED bestätigte dann den Treffer. Erstaunlicherweise haben hierbei zwei Transkriptionsfaktoren die gleichen Bindestellen.

Tabelle 8: ACTA-1-Promotor mit TFBS für die Matrixsuche [URL-10]

Menschlicher ACTA-1 Promotor
>ACTA1:chr1:225969233 [-700..100](-) [human, Homo sapiens] CTCCGGCCGCCGGTGGCCCTCTGTGCGGTGGGGGAAGGGGTCGACGTGGC 50 TCAGCTTTTTGGATTCAGGGAGCTCGGGGGTGGGAAGAGAGAAATGGAGT 100 TCCAGGGGCGTAAAGGAGAGGGAGTTCGCCTTCCTTCCCTTCCTGAGACT 150 CAGGAGTGACTGCTTCTCCAATCCTCCCAAGCCCACCACTCCACACGACT 200 CCCTCTTCCCGGTAGTCGCAAGTGGGAGTTTGGGGATCTGAGCAAAGAAC 250 CCGAAGAGGAGTTGAAATATTGGAAGTCAGCAGTCAGGCACCTTCCCGAG 300 CGCCCAGGGCGCTCAGAGTGGACATGGTTGGGGAGGCCTTTGGGACAGGT 350 GCGGTTCCCGGAGCGCAGGCGCACACATGCACCCACCGGCGAACGCGGTG 400 ACCCTCGCCCCACCCCATCCCCTCCGGCGGGCAACTGGGTCGGGTCAGGA 450 GGGGCAAACCCGCTAGGGAGACACTCCATATACGGCCCGGCCCGCGTTAC 500 CTGGGACCGGGCCAACCCGCTCCTTCTTTGGTCAACGCAGGGGACCCGGG 550 CGGGGGCCCAGGCCGCGAACCGGCCGAGGGAGGGGGCTCTAGTGCCCAAC 600 ACCCAAATATGGCTCGAGAAGGGCAGCGACATTCCTGCGGGGTGGCGCGG 650 AGGGAATGCCCGCGGGCTATATAAAACCTGAGCAGAGGGACAAGCGGCCA 700 CCGCAGCGGACAGCGCCAAGTGAAGCCTCGCTTCCCCTCCGCGGCGACCA 750 GGGCCCGAGCCGAGAGTAGCAGTTGTAGCTACCCGCCCAGGTAGGGCAGG 800 A 801

TATA-Box S = 8,07	
AGL3 S = 6,03 / 9,39	
SRF S = 5,50 / 9,40	

5.3 CSF-2 Promotor

Die Gen ID CSF-2 steht für "Colony Stimulating Factor 2". Dieses Gen codiert ein Zytokin-Protein, dass die Produktion, Differenzierung und Funktion von Granulozyten und Makrophagen steuert. Die aktive Form des Proteins, kommt im extrazellulären Bereich als Homodimer vor. Das Gen befindet sich auf Chromosom 5 in einem Gen-Cluster. Das sind Gene die zu einer Gruppe zusammengefasst sind, weil sie ähnliche Funktionen erfüllen. [URL-7]

Eine phylogenetische Footprintanalyse ergab die in Abbildung 18 dargestellten Ergebnisse. Es besteht eine hohe Übereinstimmung der Mensch-Maus Sequenzen ca. 200 bp vor Transkriptionsstart. Es wurden insgesamt 17 TFBS in einem sehr engen Sequenzfenster gefunden.

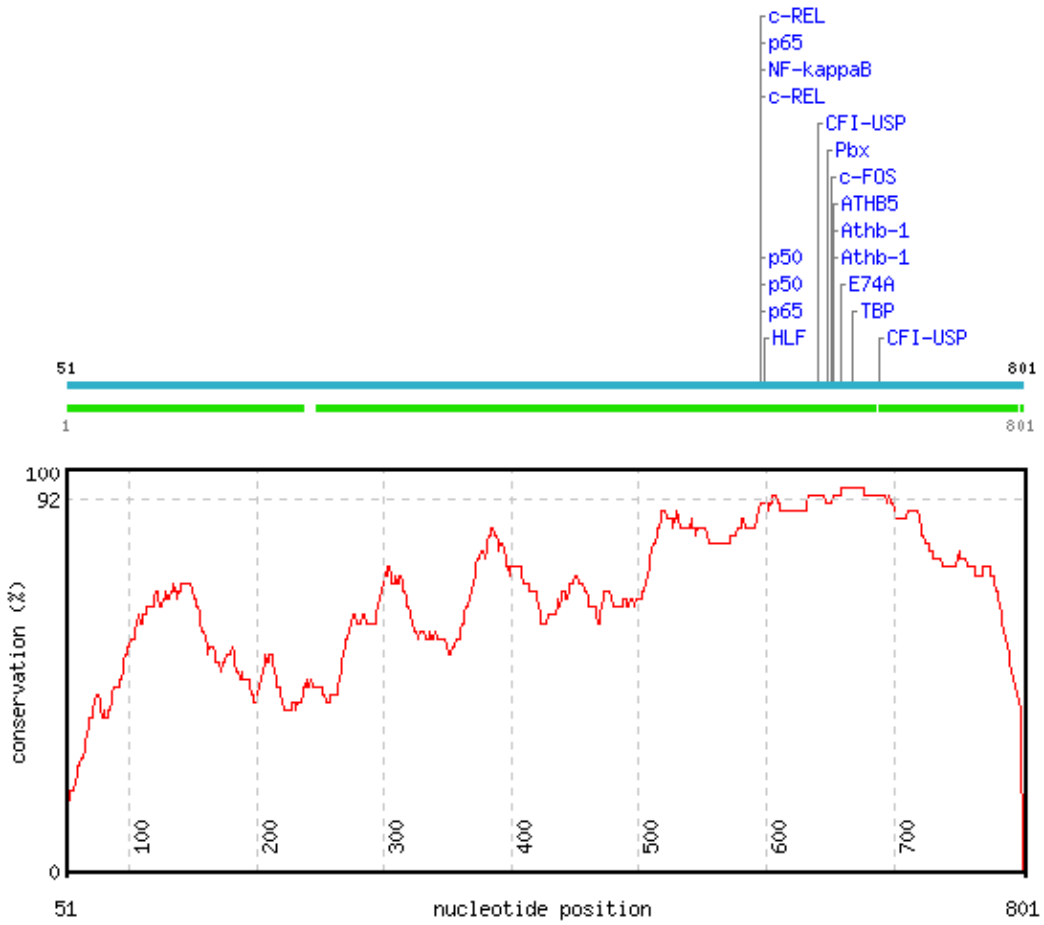
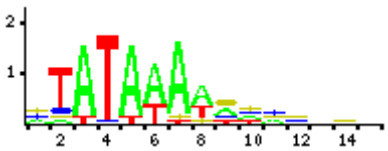
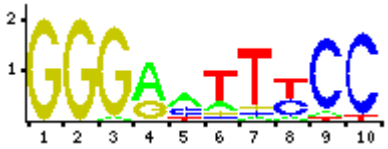
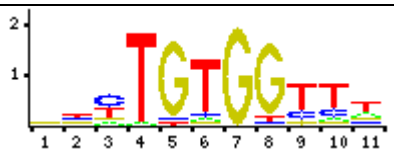


Abbildung 18: Phylogenetisches Footprinting des CSF-2 Promotors [URL-9]

Eine anschließende Matrix-Suche zeigt die Ergebnisse in Tabelle 9.

Tabelle 9: CSF-2 Promotor mit TFBS für die Matrixsuche [URL-10]

Menschlicher CSF-2 Promotor
>CSF2:chr5:131440347 [-700..100](+) [human, Homo sapiens] TCAGGAAGGTGGCTCCAGAGCCAAGAGTCAGACTCTGGGTCCCGACTTGA 50 CCCAGCCACACCCCCTCTGAAGCTTGCTGAGAGTGGCTGCAGTCTCGCTG 100 CTGGATGTGCACATGGTGGTCATTCCCTCTGCTCACAGGGGCAGGGGTCC 150 CCCCTTACTGGACTGAGGTTGCCCCCTGCTCCAGGTCCTGGGTGGGAGCC 200 CATGTGAACTGTCAGTGGGGCAGGTCTGTGAGAGCTCCCCTCACACTCAA 250 GTCTCTCACAGTGGCCAGAGAAGAGGAAGGCTGGAGTCAGAATGAGGCAC 300 CAGGGCGGGCATAGCCTGCCCAAAGGCCCCTGGGATTACAGGCAGGATGG 350 GGAGCCCTATCTAAGTGTCTCCCACGCCCCACCCCAGCCATTCCAGGCCA 400 GGAAGTCCAAACTGTGCCCCTCAGAGGGAGGGGGCAGCCTCAGGCCCATT 450 CAGACTGCCCAGGGAGGGCTGGAGAGCCCTCAGGAAGGCGGGTGGGTGGG 500 CTGTCGGTTCTTGGAAAGGTTCATTAATGAAAACCCCCAAGCCTGACCAC 550 CTAGGGAAAAGGCTCACCGTTCCCATGTGTGGCTGATAAGGGCCAGGAGA 600 TTCCACAGTTCAGGTAGTTCCCCCGCCTCCCTGGCATTTTGTGGTCACCA 650 TTAATCATTTCCTCTGTGTATTTAAGAGCTCTTTTGCCAGTGAGCCCAGT 700 ACACAGAGAGAAAGGCTAAAGTTCTCTGGAGGATGTGGCTGCAGAGCCTG 750 CTGCTCTTGGGCACTGTGGCCTGCAGCATCTCTGCACCCGCCCGCTCGCC 800 C

TATA-Box S = 5,19	
NF-kappaB S = 3,39	
AML-1 S = 9,91	

Die mit blauem Hintergrund markierte Sequenzen waren falsch-positive Hits für den NF-kappaB Transkriptionsfaktor. Alle Hits dieses Faktors besitzen ungefähr einen Score von 3,5. Jedoch war nur einer der Treffer ein echt-positiver (Sequenz hellblau).

6 Diskussion

6.1 IL10 - Promotor

Mithilfe der Matrix-Suche konnten 4 TFBS gefunden werden, die auch experimentell Nachgewiesen wurden. Jedoch gibt es laut Abbildung 16 insgesamt zehn TFBS. Somit wurden sechs überhaupt nicht identifiziert.

Das phylogenetische Footprinting erzielte keinen einzigen echt positiven Treffer. Die Erklärung hierfür könnte sein, dass falsche Eingabewerte wie zum Beispiel der Cutoff, die Fenstergröße oder der "TF Score Treshold" verwendet wurden. Das phylogenetische Footprinting dieses Promotors ist im Anhang zu finden.

6.2 ACTA1 - Promotor

Die TFBS, die beim Phylogenetischen Footprinting ermittelt wurden hatten einen sehr hohen Score und trafen beide auf dieselben Bindestellen zu. Die SABiosciences-Datenbank bestätigt die Bindestelle für SRF, jedoch nicht für AGL3. Anhand der beiden Sequenzlogos aus Tabelle 8 erkennt man das beide Faktoren eine ähnliche Konsensussequenz als Bindestelle besitzen und diese Tatsache auf den hohen Score für beide Sequenzen schließen lässt. Der Transkriptionsfaktor AGL-3 wurde anhand der Datenbank-Analyse als falsch-positiver Treffer bewertet.

6.3 CSF2 - Promotor

Die Suche von TFBS auf der Promotorsequenz mit ConSite ergab nur wenige richtige Treffer. Hervorzuheben ist, dass NFkappa-B, ein wichtiger Transkriptionsfaktor bei der Regulation des CSF2-Gens als echt positiver Treffer gefunden wurde. Außerdem wurde noch das TATA-Bindeprotein als richtig erkannt. Die restlichen Hits konnten nur als falsch-positive Treffer charakterisiert werden, da sie auf keiner anderen Datenbank bestätigt werden konnten. Die Matrixsuche auf TRED ergab für den Transkriptionsfaktor AML-1 einen echt positiven Treffer, der auch mit anderen Datenbanken bestätigt werden konnte. Mit einem Score von 9,91 ist dies eine sichere Bindestelle.

7 Ausblick

Das Verfahren des phylogenetischen Footprintings mit einer anschließenden Matrixsuche eignete sich sehr gut um potentielle TFBS vorherzusagen. Der Nachteil daran ist jedoch, dass man auch viele falsch positive Ergebnisse erhält. Also werden TFBS angezeigt, die eigentlich gar keine sind. Manche TFBS werden auch gar nicht erst gefunden. Der Grund dafür war nicht sofort ersichtlich, jedoch gab ein Paper der Universitätsmedizin Göttingen Aufschluss darüber. Es gibt Transkriptionsfaktoren, die die Bindung eines weiteren TF begünstigen. Dies ist zum Beispiel bei NF-ATp der Fall. Besitzt dieser TF eine Bindestelle, wie in Abbildung 19 zu sehen, dann begünstigt dies die Anbindung des Transkriptionsfaktors AP-1 obwohl dieser eine von der tatsächlichen bindestelle abweichende DNA-Sequenz bindet.

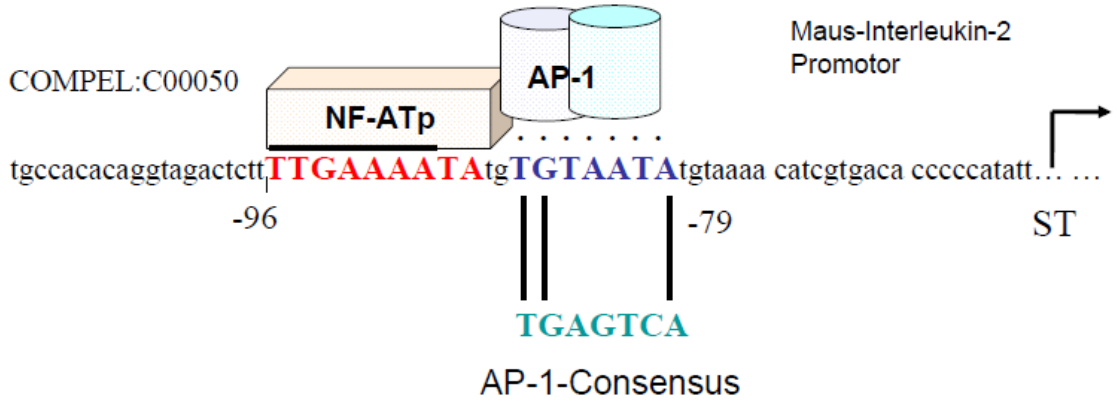


Abbildung 19: Interaktion zwischen Transkriptionsfaktoren [URL-12]

Solche und ähnliche Interaktionen zwischen Transkriptionsfaktoren machen das phylogenetische Footprinting für Spezialfälle wie diesen nahezu unbrauchbar. Man kann also sagen, dass man mit dem Verfahren grobe Aussagen über Bindestellen von Transkriptionsfaktoren machen kann, diese jedoch unbedingt experimentell bestätigen sollte.

8 Zusammenfassung

In meiner Bachelorarbeit konnte ich zeigen, dass das Phylogenetische Footprinting durchaus eine Methode ist um potentielle Bindestellen für Transkriptionsfaktoren vorherzusagen. Dieses Verfahren wurde in der Zeit meiner Arbeit für einige ausgewählte Promotorregionen von Genen durchgeführt. Es ist anzumerken, dass durch das phylogenetische Footprinting sämtliche weitere Promotorregionen analysiert und dies somit im großen Maßstab angewandt werden kann. Zum Ende hin gab es einige Schwierigkeiten, da jedes einzelne Gen durch sehr komplizierte Vorgänge reguliert wird. Oftmals ist es auch so, dass ein Gen bis zu 50 oder mehr Bindestellen für TF aufweist und es sich als sehr schwierig zeigte, den Überblick zu behalten. Eine weitere Komplikation stellte sich bei der Tatsache heraus, dass in unterschiedlichen Geweben oder Zellen auch unterschiedliche Bedingungen für TF herrschen. Man kann somit kein einheitliches Bild für die Transkriptionsfaktoren, die ein Gen regulieren erstellen. Man müsste also für jeden Gewebe bzw. Zelltyp ein neues Schema anfertigen. Zudem ist die gegenseitige Interaktion von Transkriptionsfaktoren anzumerken (Abbildung 19), die es für Suchmatrizen fast unmöglich macht für solche Fälle echt positive Treffer zu landen. Einfachere Regionen, wie die TATA-Box, die in fast jedem eukaryontischen Promotor zu finden ist, sind dagegen leicht zu identifizieren. Schlussendlich wurde das Ziel sämtliche Promotorregionen auf TF zu untersuchen etwas zu global betrachtet. In der Biologie der Organismen muss man etwas genauer sein.

Eine Verbesserung für die Suchergebnisse beim Phylogenetischen Footprinting könnte man dadurch erreichen, zusätzlich regulatorische Netzwerke einzubeziehen, damit Interaktionen zwischen Transkriptionsfaktoren mit bei der Suche berücksichtigt werden.

9 Summary

In my bachelor thesis I have shown that phylogenetic footprinting is definitely a method to predict potential binding sites for transcription factors. This procedure was performed in the time of my work for a few selected promoter regions of genes. It should be noted that by phylogenetic footprinting any further promoter regions and this can be applied on a larger scale. Toward the end, there were some difficulties, because each gene is regulated by highly complex processes. Often it is also the case that a gene got up to 50 or more binding sites for transcription factors and it was hard to keep overview. A further complication was found in the fact that in different tissues or cells are also different conditions for transcription factors. Thus you can't create a uniform picture for transcription factors which regulate a gene. You would have to customize a new pattern for each tissue or cell type. In addition, the mutual interaction should be noted (Figure 19), which makes it almost impossible for scoring matrices to land true positive results for such cases. Easier regions such as the TATA-box, which is found in almost all eukaryotic promoters are easy to detect. Finally, the destination to examine all the promoter regions for transcription factors was considered too global. In the biology of species you have to be more accurate. An improvement of search results through phylogenetic footprinting could be achieved by including regulatory networks, thus interacts between transcription factors are considered.

Literaturverzeichnis

[1] Sauer, Tilman 2006 *Evaluierung des phylogenetischen Footprintings zur verbesserten Vorhersage von Transkriptionsfaktor-Bindestellen*. 148 Seiten, Göttingen, Georg - August Universität Göttingen, Mathematisch - Naturwissenschaftliche Fakultät

[2] Renneberg, Reinhard; Süßbier, Darja (2010): *Biotechnologie für Einsteiger*. 3. Auflage: Spektrum Akademischer Verlag

[3] Clark, David P. (2006): *Molecular Biology Understanding the Genetic Revolution*. 1. Auflage: Spektrum Akademischer Verlag

[4] Alberts, B.; Bray, D.; Hopkin, K.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. (2005): Lehrbuch der Molekularen Zellbiologie. 3. Auflage: WILEY-VCH Verlag

[URL-2] (20/07/2013) Reusch, Sebastian: *Chemie: Gut oder Böse? Giftig oder nicht?* URL: http://www.scilogs.de/wblogs/blog/enkapsis/aufklarung-irrtumer/2011-04-08/chemie-gut-oder-b-se-giftig-oder-nicht

[URL-3] (04/08/2013) Brüning, Anne: *Das Buch des Lebens für die Ratte*. URL: http://www.berliner-zeitung.de/archiv/mit-dem-erbgut-der-laborratte-haben-forscher-das-dritte-saeugetiergenom-entziffert--es-soll-helfen--menschliche-krankheiten-zu-verstehen-das-buch-des-lebens-fuer-die-ratte,10810590,10164928.html

[URL-4] (24/06/13) Wasserman, Wyeth: *Predicting transcriptional regulation from gene lists.* URL: http://bioinformatics.ca//files/GeneLists_Day2-Module3.pdf

[URL-5] (06/08/2013) EMBL-EBI: Pairwise Sequence Alignment (Nucleotide) URL: http://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html

[URL-6] (24/06/2013) kein Autor: *EPD - Eukaryotic Promoter Database*. URL: http://epd.vital-it.ch/

[URL-7] (02/05/2013) kein Autor: *GeneCards The Human Gene Compendium*. URL: http://www.genecards.org/

[URL-8] (02/08/2013) kein Autor: *UCSC Genome Bioinformatics*. URL: http://genome.ucsc.edu/index.html

[URL-9] (15/08/2013) kein Autor: *ConSite.*
URL: http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite

[URL-10] (19/08/2013) kein Autor: *JASPAR*.
URL: http://jaspar.genereg.net/cgi-bin/jaspar_db.pl

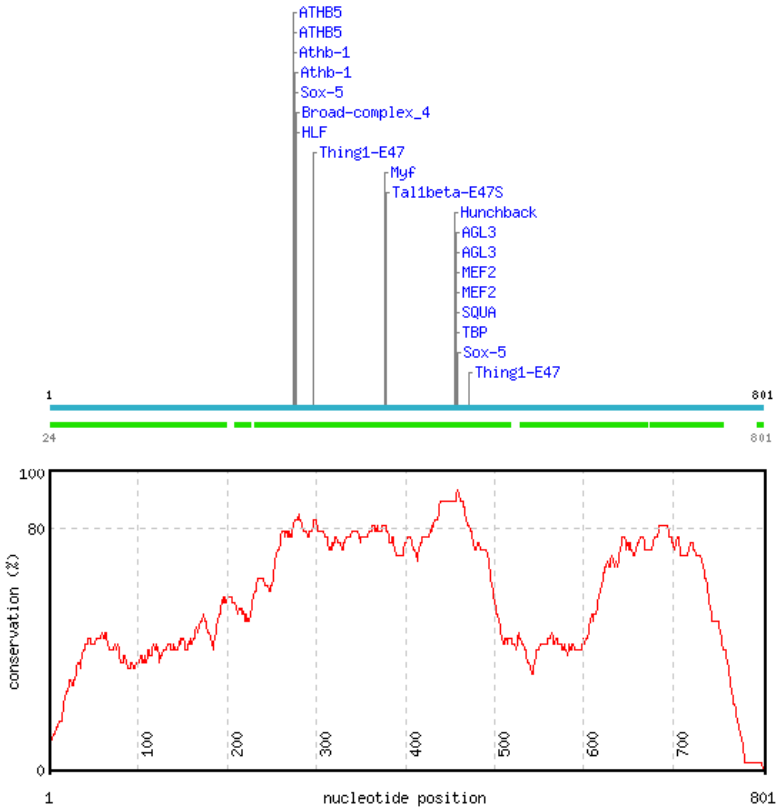
[URL-11] (19/08/2013) Saraiva, Margarida: *Molecular regulation of interleukin-10 expression*. URL: http://www.nature.com/nri/journal/v10/n3/fig_tab/nri2711_F3.html

[URL-12] (19/08/2013) Morgenstern, Burkhard: *Grundlagen der Bioinformatik*. URL: http://www.bioinf.med.uni-goettingen.de/fileadmin/upload/teaching/bio/bioinf_grund/2008/biologie_grund_080519.pdf

[URL-13] (02/08/2013) Zhang, Michael: *Transcriptional Regulatory Element Database*. URL: http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=searchOrthForm

Anhang

Phylogenetisches Footprinting IL10 - Promotor



Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, den 26.08.2013

_____

Toni Trodler